Title: Modeling for Prediction –
Linear Regression with Excel, Minitab, Fathom and the TI-83

Brief Overview:

In this lesson section, the class is going to be exploring data through linear regression while making use of four technology packages. Students will organize, represent, and interpret data on given sets of data. They will make predictions and explore the appropriateness of models connected with the data. Students will calculate descriptive statistics, draw plots, and obtain different regression models using statistical software. Within this packet, there are step-by-step instructions for the use of four statistical technologies; Excel, Minitab, TI-83/TI-83+, and Fathom.

NCTM Content Standard/National Science Education Standard:

The AP Statistics curriculum is listed at www.collegeboard.com

Grade/Level:

10-12; AP Statistics classes

Duration/Length:

Approximately 3-5 (three) 60 minute Classes/Blocks .

Student Outcomes:

Students will:

- Learn to create lists in the TI-83, and worksheets in Excel and in Minitab/Fathom.
- Calculate descriptive statistics for data.
- Learn to create scatter plots with data.
- Learn to add summary (mean) coordinates to a scatter plot.
- Learn to generate a least squares regression equation.
- Learn to plot that equation against observed data values.
- Learn the difference between interpolation and extrapolation, the pitfalls of each, and how to generate interpolated and extrapolated values.
- Learn to generate and interpret a residuals plot.
- Compare and contrast different technologies and their various output capabilities.

Materials and Resources:

- TI-83 or TI-83 Plus calculator.
- A computer with Excel 2000 or above, Minitab 12 or above, and/or Fathom
- Teacher notes
- Student worksheets
- Resource sheets
- Teacher answer keys

Development/Procedures:

Lesson 1 - Exploring Regression with the TI-83

Launch – Explore the TI-83's list making capabilities, the creation and storage of variables.

Teacher Facilitation – Presentation of the concept of a mean values coordinate and the idea of linear regression. Work with students to develop their understanding of the concept through the use of technology and worksheet manipulatives.

Student Application – Allow students to build knowledge in depth by using different technologies that solve the same problem.
Allow students to make professional presentations of their discoveries and insights.

Embedded Assessment – Utilize problems and worksheets to determine each student's progress toward understanding of the concept.
Build upon this understanding by having them present their material or to publish it in a report.

Reteaching/Extension –

- For those who have not completely understood the lesson, review what is needed.
- For those who have understood the lesson, take them to the next step in development of the concept.

Lesson 2 - Exploring Regression with Excel using the Data Analysis ToolPak

Launch – Explore Excel's worksheet making capabilities, the creation and storage of variables.

Teacher Facilitation – Presentation of the concept of a mean values coordinate and the idea of linear regression. Work with students to develop their understanding of the concept through the use of technology and worksheet manipulatives.

Student Application – Allow students to build knowledge in depth by using different technologies that solve the same problem.
Allow students to make professional presentations of their discoveries and insights.

Embedded Assessment – Utilize problems and worksheets to determine each student's progress toward understanding of the concept. Build upon this understanding by having them present their material or to publish it in a report.

Reteaching/Extension –

- For those who have not completely understood the lesson, review what is needed.
- For those who have understood the lesson, take them to the next step in development of the concept.

Lesson 3 - Exploring Regression with Minitab Data Analysis Software

Launch – Explore Minitab's worksheet capabilities, the creation and storage of variables.

Teacher Facilitation – Presentation of the concept of a mean values coordinate and the idea of linear regression. Work with students to develop their understanding of the concept through the use of technology and worksheet manipulatives.

Student Application – Allow students to build knowledge in depth by using different technologies that solve the same problem. Allow students to make professional presentations of their discoveries and insights.

Embedded Assessment – Utilize problems and worksheets to determine each student's progress toward understanding of the concept. Build upon this understanding by having them present their material or to publish it in a report.

Reteaching/Extension –

- For those who have not completely understood the lesson, review what is needed.
- For those who have understood the lesson, take them to the next step in development of the concept.

Lesson 4 - Exploring Regression with Fathom Data Analysis Software

Launch – Explore Fathom's worksheet capabilities, the creation and storage of variables.

Teacher Facilitation – Presentation of the concept of a mean values coordinate and the idea of linear regression. Work with students to develop their understanding of the concept through the use of technology and worksheet manipulatives.

Student Application – Allow students to build knowledge in depth by using different technologies that solve the same problem. Allow students to make professional presentations of their discoveries and insights.

Embedded Assessment – Utilize problems and worksheets to determine each student's progress toward understanding of the concept. Build upon this understanding by having them present their material or to publish it in a report.

Reteaching/Extension –

- For those who have not completely understood the lesson, review what is needed.
- For those who have understood the lesson, take them to the next step in development of the concept.

Authors:

Name		Duke M Writer Jr
School		Potomac Falls High School
County		Loudoun County
State		Virginia

Name		William Marbury
School		Sandy Spring Friends School
County		Montgomery County
State		Maryland

Introduction

Exploring Linear Regression Analysis with Excel, Minitab, the TI-83 and Fathom

The Sample Problem:

The following is the sample problem, which will be used for each of the technology instruction sets below.

We want to explore whether or not a relationship exists between the number of sales clerks on duty in a retail store (response variable - x) and the amount of losses due to shrinkage (i.e. shoplifting, damaged merchandise and other material loss).

Data:

X — Response Variable = Staffing Level (Staff) = the number of sales clerks on duty
Y — Explanatory Variable = Shrinkage (Cost) = the dollar ($) amount of shrinkage, in 100's of dollars

Staff	Cost
10	19
12	15
11	20
15	9
9	25
13	12
8	31

Part I. Using the TI-83 in Linear Regression Analysis

Entering Data and Creating a Scatterplot in the TI-83

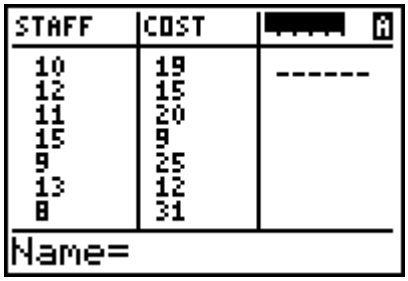
1. First we need to setup the List Editor. We will define our explanatory variable (the number of clerks on duty) as "STAFF" and our response variable (shrinkage in hundreds of dollars) as "COST":

   STAT / <5:SetUpEditor> / 2ND / ALPHA / **STAFF** / ALPHA / , / 2ND / ALPHA / **COST**

   Make sure that you type in the comma between **STAFF** and **COST**.

2. Now, we will go to the list editor to enter our data:
   STAT / <1:Edit> / ENTER

3. You're now ready to enter in the data. Your screen should look like the figure on the right when you're finished.



4. To make a scatter plot from our data, do the following:
   2ND / Y= / <1: Plot1 … > / ENTER

   This gets you into the Plot Editor for Plot1.

5. You want to activate Plot1 by highlighting the ON button and pressing ENTER.

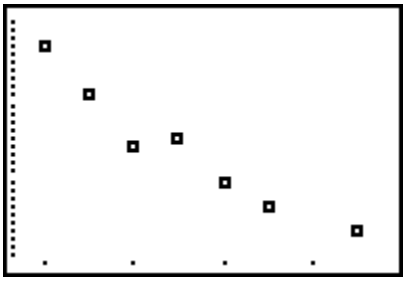
   Under < Type: > highlight the scatterplot icon and press ENTER.

   Under < XList: > press 2ND / STAT and use the toggle down key (⇓) to find and highlight the variable **STAFF**. Then press ENTER.

   Under < YList: > press 2ND / STAT and use the toggle down key (⇓) to find and highlight the variable **COST**. Then press ENTER. Now your screen should look like this:

6. Finally, press the blue ZOOM key, then press < 9:ZoomStat > This will automatically set the window size and create a scatterplot of your data.

   The scatter plot screen should look like this:



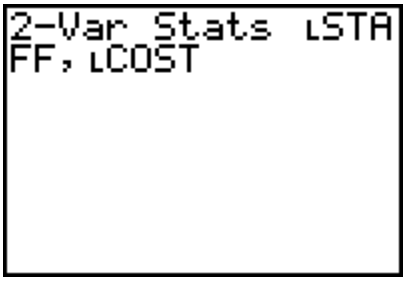
Creating and Plotting a Mean Values Coordinate in the TI-83

1. Once we have created our Scatter Plot, we might want to add an additional coordinate, viz. the mean values of x and y or $(\bar{x}, \bar{y})$. Use **2-Variable Stats** summary:

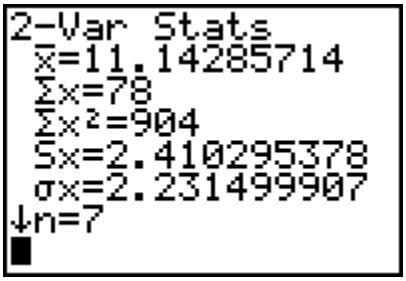
   STAT / < 2: 2-Var Stats > / ENTER

   Now, you will need to define your variables:

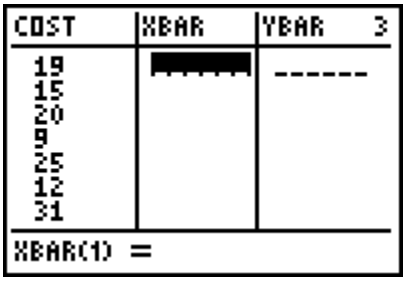
   2ND / STAT and use the toggle down key (⇓) to find and highlight the variable **STAFF**. Type , / 2ND / STAT and use the toggle down key to find and highlight the variable **COST**.



2. Press ENTER to see the 2-Var Stats Summary. The first statistic is shown is $\bar{x}$ = 11.14285714. If you use the toggle down key (⇓), you will be able to find that $\bar{y}$ = 18.71428571. This means that the mean value coordinate $(\bar{x}, \bar{y})$, in this case, is the point (11.14, 18.71). Thus, the average number of clerks was 11.14, while the average dollar amount of shrinkage was $1871.



3. We can now put this mean value coordinate $(\bar{x}, \bar{y})$ to some use. First we want to plot these values into our scatter plot. Press STAT / < 1: Edit … > / ENTER. You are now back in the List Editor. Toggle over to the top of the empty column to the right of the COST column, as shown: Now, enter a new variable name: XBAR . Next, toggle to the right one column and enter a second new variable name: YBAR , as shown:

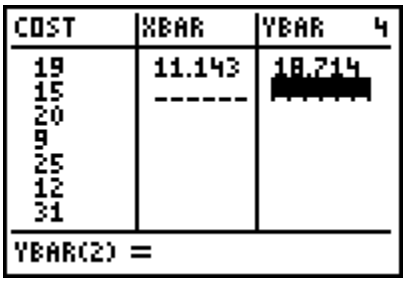


Make sure that the cursor is highlighted as shown in the first row of the **XBAR** column. Now fill this in with the actual value for $\bar{x}$.

Press VARS / ⇓ / < 5: Statistics … > / ENTER / ⇓ / < 2: $\bar{x}$ > / ENTER / ENTER

Now make sure that the cursor is highlighted as shown in the first row of the **YBAR** column.

Press VARS / ⇓ / < 5: Statistics … > / ENTER / ⇓ / < 2: $\bar{y}$ > / ENTER / ENTER.



4. Now, we want to plot the coordinate $(\bar{x}, \bar{y})$.

   Press 2ND / Y= to re-enter the Plot Editor.

   Toggle down (⇓) and choose < 2: Plot 2 … > / ENTER.

   You want to activate Plot2 by highlighting the ON button and pressing ENTER.

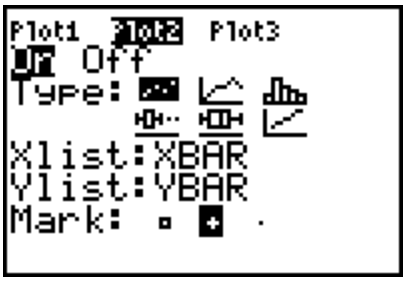
   Under < Type: > highlight the scatterplot icon and press ENTER.

   Under < XList: > press 2ND / STAT and use the toggle down key (⇓) to find and highlight the variable **XBAR**.

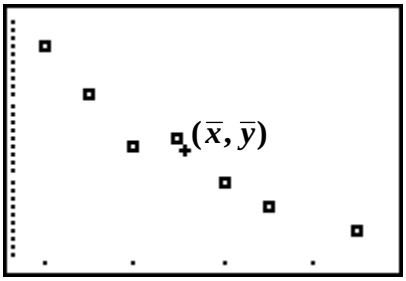
   Then press ENTER.

   Under < YList: > press 2ND / STAT / toggle down key (⇓) to find and highlight the variable **YBAR**. Press ENTER.

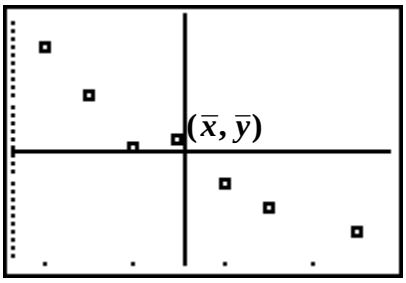
   Mark: should be a cross +



5. Make sure that both Plot 1 and Plot 2 are on by pressing Y=. Both Plot 1 and Plot 2 should be highlighted at the top of the menu. Now press ZOOM / ⇓ / < 9: ZoomStat > / ENTER. Your scatter plot now will show your data and the mean values coordinate.



   Notice that this coordinate breaks the scatter plot into four regions. In our problem we can see that four observed values lie <u>above and to the left</u> of the mean values $(\bar{x}, \bar{y})$, and that the remainder of the data lies <u>below and to the right</u> of the mean values.



Creating a Least-Squares Regression Line in the TI-83

1. To generate our linear regression line and then to plot it with our data, we start with:

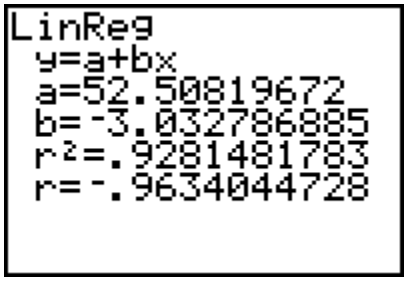
   STAT / ⇓ / < CALC > / < 8: LinReg (a +bx) >

   Now you will need to identify your explanatory (x) variable:

   2ND / STAT and use the toggle down key (⇓) to find and highlight the variable **STAFF**. Now type a COMMA

   Then press 2ND / STAT and use the toggle down key (⇓) to find and highlight the variable **COST**. Now type a COMMA

   Next, in order to store the regression equation into Y_1, press VARS / ⇒ / < Y-VARS > / < 1: Function > / < 1: Y_1 >



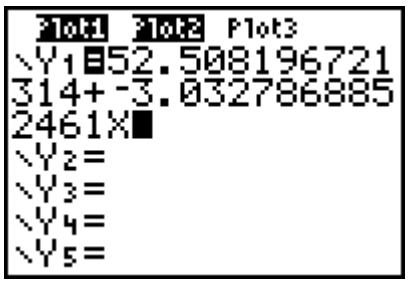
2. Press ENTER to get the LinReg equation:

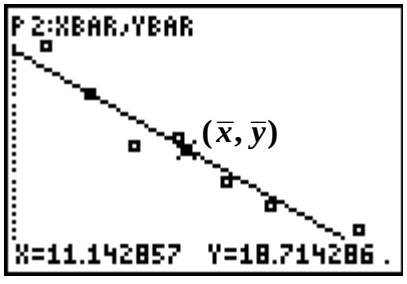
   If your correlation coefficient and coefficient of determinantion do not appear, do the following:

   2ND / 0 / x^{-1} and use the toggle down key (⇓) to find < DiagnosticOn >. Press ENTER / ENTER.

3. Now we can look at the regression equation in Y_1.

   Press the blue key Y= and press TRACE.

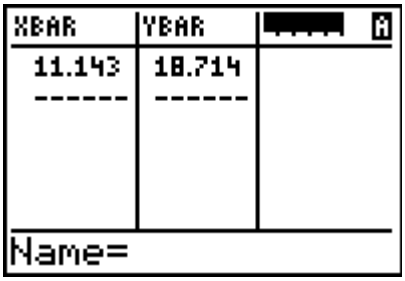




Creating a List of Residuals and Plotting a Residual Plot

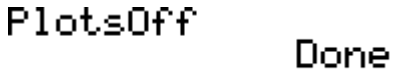
1. After taking the Linear Regression of a set of data (< LinReg >), a list of the residuals will automatically made and stored in the calculator. In order to place that list in the List Editor, press STAT / < 1:Edit… >.

   Toggle to the List Names and toggle across to an open list.

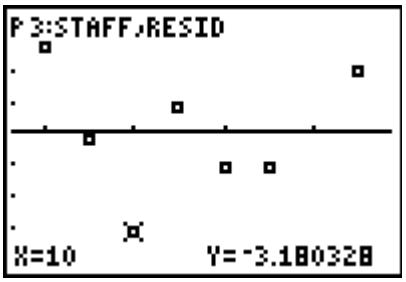


2. Press 2ND / STAT, toggle down to **RESID**, and press ENTER / ENTER.

3. To create the scatter plot of the residuals versus the number of clerks, turn off all plots first.



   Press 2ND / Y= for the StatPlot window. Choose Plot3. Turn the plot ON, chose the first type of plot (scatter). Define the Xlist as **STAFF** and the Ylist as **RESID**. Choose a mark and press ZOOM / < 9:ZoomStat > and press TRACE

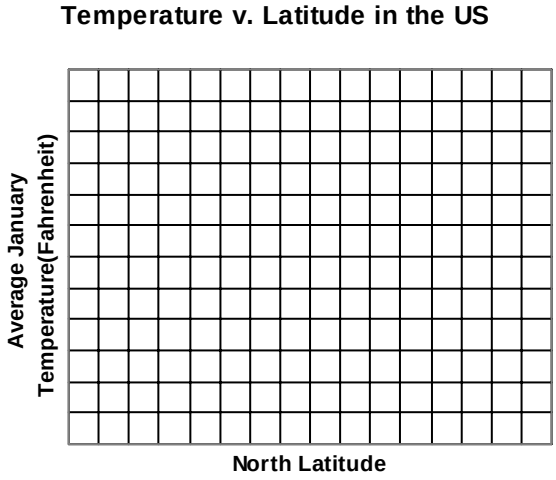


Linear Regression Worksheet
Average January Temperature

Name: _____ Teacher: _____ Date: _____

A) Draw a scatterplot of these data.

City	JanTemp	Lat
Albuquerque, NM	24	35.1
Amarillo, TX	24	35.6
Baltimore, MD	25	39.7
Boise, ID	22	43.7
Boston, MA	23	42.7
Cincinnati, OH	26	39.2
Concord, NH	11	43.5
Denver, CO	15	40.7
Detroit, MI	21	43.1
Houston, TX	44	30.1
Indianapolis, IN	21	39.8
Key West, FL	65	25
Los Angeles, CA	47	34.3
Madison, WI	9	43.4
Minneapolis, MN	2	45.9
Montgomery, AL	38	32.9
New York, NY	27	40.8
Philadelphia, PA	24	40.9
Phoenix, AZ	35	33.6
Portland, ME	12	44.2
San Francisco, CA	42	38.4
Seattle, WA	33	48.1
Spokane, WA	19	48.1
Washington, DC	30	39.7

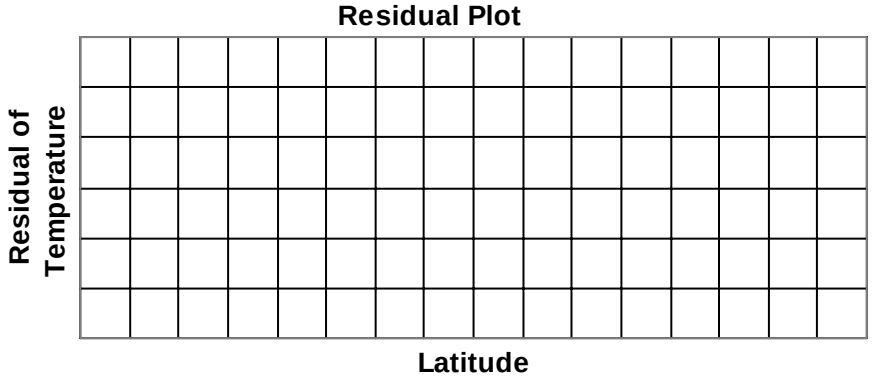


B) Find the regression equation and correlation for the Temperature and Latitude data.

C) Fill out the following table.

City	January Temp (y)	Latitude	Predicted Latitude $(\hat{y})$	Residual Value $(y-\hat{y})$
Key West, FL	65	25		
Houston, TX	44	30.1		
Montgomery, AL	38	32.9		
Phoenix, AZ	35	33.6		
Los Angeles, CA	47	34.3		
Albuquerque, NM	24	35.1		
Amarillo, TX	24	35.6		
San Francisco, CA	42	38.4		
Cincinnati, OH	26	39.2		
Washington, DC	30	39.7		
Baltimore, MD	25	39.7		
Indianapolis, IN	21	39.8		
Denver, CO	15	40.7		
New York, NY	27	40.8		
Philadelphia, PA	24	40.9		
Boston, MA	23	42.7		
Detroit, MI	21	43.1		
Madison, WI	9	43.4		
Concord, NH	11	43.5		
Boise, ID	22	43.7		
Portland, ME	12	44.2		
Minneapolis, MN	2	45.9		
Seattle, WA	33	48.1		
Spokane, WA	19	48.1		

D) Determine whether or not this model is appropriate for the given data. (Make sure that you draw a residual plot of the data.)



E) Do you need to change your model to make it a better fit? How?

F) What would you predict to be the average January Temperature of Mobile, AL (Lat = 31.2)? How does this prediction compare to the actual observation at 44 degrees? Is this prediction good or bad?

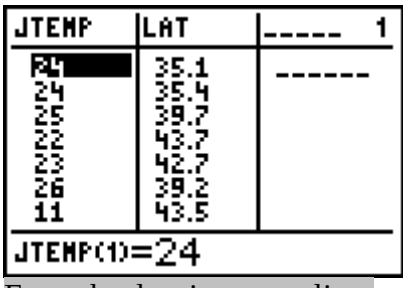
G) Can you make a prediction of the average January Temperature on the Equator (Lat = 0)? What problems do you have with this prediction?

Linear Regression Worksheet
Average January Temperature

Name: KEY Teacher: KEY Date: KEY

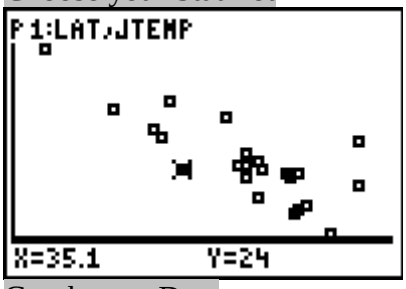
A) Draw a scatterplot of these data.

City	JanTemp	Lat
Albuquerque, NM	24	35.1
Amarillo, TX	24	35.6
Baltimore, MD	25	39.7
Boise, ID	22	43.7
Boston, MA	23	42.7
Cincinnati, OH	26	39.2
Concord, NH	11	43.5
Denver, CO	15	40.7
Detroit, MI	21	43.1
Houston, TX	44	30.1
Indianapolis, IN	21	39.8
Key West, FL	65	25
Los Angeles, CA	47	34.3
Madison, WI	9	43.4
Minneapolis, MN	2	45.9
Montgomery, AL	38	32.9
New York, NY	27	40.8
Philadelphia, PA	24	40.9
Phoenix, AZ	35	33.6
Portland, ME	12	44.2
San Francisco, CA	42	38.4
Seattle, WA	33	48.1
Spokane, WA	19	48.1
Washington, DC	30	39.7



Enter the data into your lists.

Choose your StatPlot



Graph your Data

B) Find the regression equation and correlation for the Temperature and Latitude data.

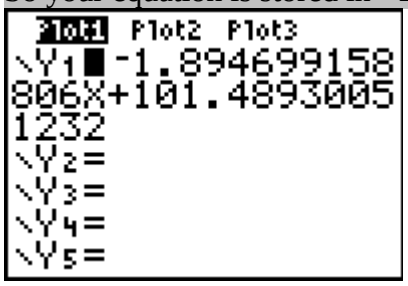
Answer: In your calculator, enter in the line

```
LinReg(ax+b) LLA
T,LJTEMP,Y1
```

You get…

```
LinReg
 y=ax+b
 a=-1.894699159
 b=101.4893005
 r²=.5998172561
 r=-.7744786996
```

So your equation is stored in "Y="

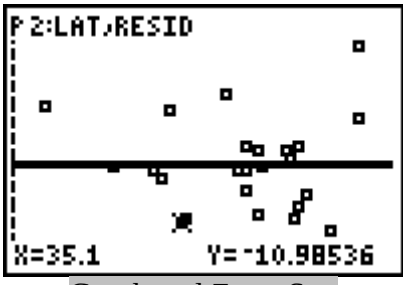


C) Fill out the following table.

City	January Temp (y)	Latitude	Predicted Temp $(\hat{y})$	Residual Value $(y-\hat{y})$
Albuquerque, NM	24	35.1	34.985	-10.99
Amarillo, TX	24	35.6	34.417	-10.42
Baltimore, MD	25	39.7	26.27	-1.27
Boise, ID	22	43.7	18.691	3.3091
Boston, MA	23	42.7	20.586	2.4144
Cincinnati, OH	26	39.2	27.217	-1.217
Concord, NH	11	43.5	19.07	-8.07
Denver, CO	15	40.7	24.375	-9.375
Detroit, MI	21	43.1	19.828	1.1722
Houston, TX	44	30.1	44.459	-.4589
Indianapolis, IN	21	39.8	26.08	-5.08
Key West, FL	65	25	54.122	10.878
Los Angeles, CA	47	34.3	36.501	10.499
Madison, WI	9	43.4	19.259	-10.26
Minneapolis, MN	2	45.9	14.523	-12.52
Montgomery, AL	38	32.9	39.154	-1.154
New York, NY	27	40.8	24.186	2.8144
Philadelphia, PA	24	40.9	23.996	.0039
Phoenix, AZ	35	33.6	37.827	-2.827
Portland, ME	12	44.2	17.744	-5.744
San Francisco, CA	42	38.4	28.733	13.267
Seattle, WA	33	48.1	10.354	22.646
Spokane, WA	19	48.1	10.354	8.6457
Washington, DC	30	39.7	26.27	3.7303

D) Determine whether or not this model is appropriate for the given data. (Make sure that you draw a residual plot of the data.)**Make sure to first turn off all your graphs.**

Enter the lists into Plot2



Graph and ZoomStat

E) Do you need to change your model to make it a better fit? How?
Answer: Looking at the residual plot, the linear model seems to be appropriate for the given data set. There don't seem to be any outliers, so there is no need to change this model.

F) What would you predict to be the average January Temperature of Mobile, AL (Lat = 31.2)? How does this prediction compare to the actual observation of 44 degrees? Is the prediction good or bad?
Answer: The prediction for $x = 31.2$ is 42.375, which has a residual of 1.625. This is a very good prediction shown by the very small residual.

G) Can you make a prediction of the average January Temperature on the Equator (Lat = 0)? What problems do you have with this prediction?
Answer: The prediction for $x = 0$ is 101.49 degrees. This seems to be a reasonable prediction, but it is not trustworthy to extrapolate outside of the data range.

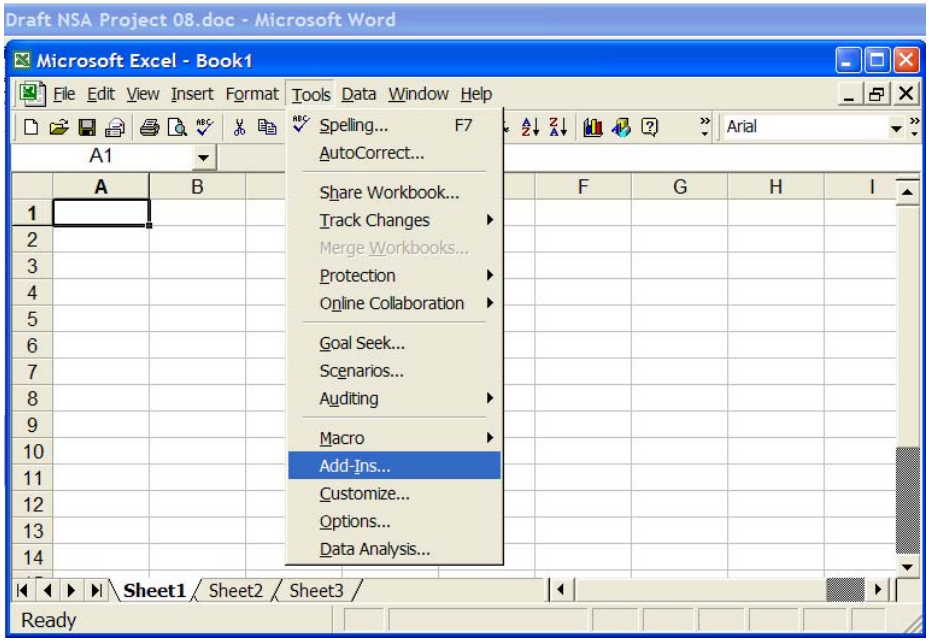
Part II. Using Excel in Linear Regression Analysis

Sources Consulted: [Excel Guide for Understandable Stats, Brase and Brase, Houghton Mifflin, 2003; Excel Manual to accompany Weiss's Introductory Statistics. Zehna, Addison Wesley, 1999]

Installing the Data Analysis ToolPak for Excel

1. In order to begin doing data analysis in **Excel** you will need to install the **Analysis ToolPak Add-In** which comes with the more recent versions of Excel. You may need your Office Installation CD.

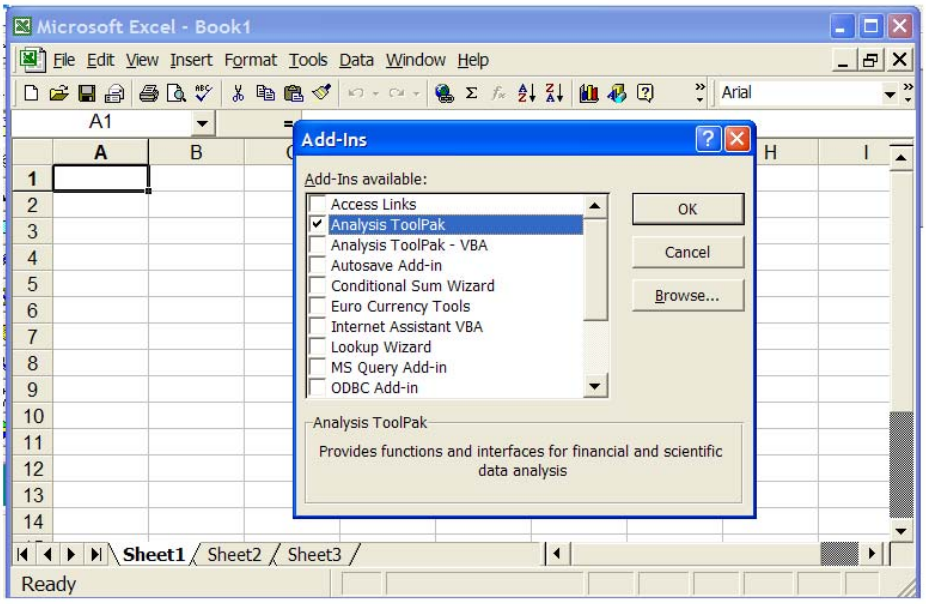
2. Click on the **Tools** menu and find **Add-Ins**, as shown:



3. Find the box for **Analysis ToolPak** and mark it with a check, as shown:

   Windows may ask you to insert your MS Office Installation CD. You may need the assistance of your System or IT Administrator.

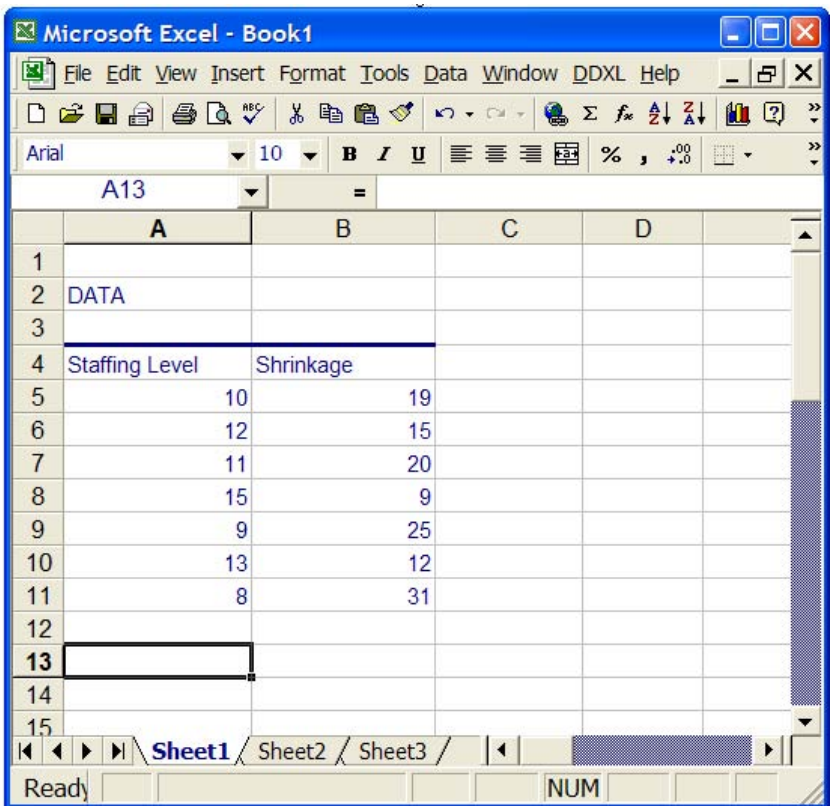
   Press OK and you are ready to go.



Entering Data into Excel Worksheet

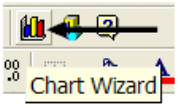
1. This step assumes that you are somewhat familiar with entering data into a table or a spreadsheet. We will use column A for the explanatory variable data, and column B for the response variable data. It will be useful to have a descriptive label for the columns.

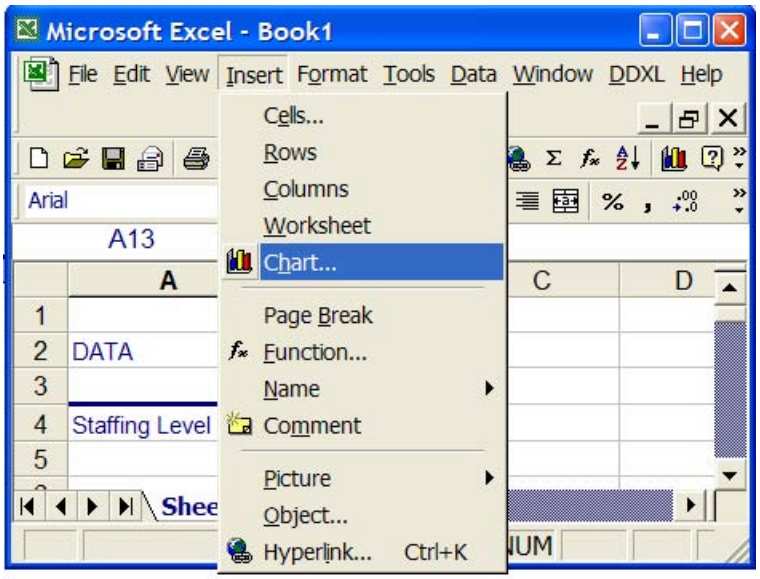
2. Enter the sample data as shown in the screen shot on the right.



Creating a Scatter Plot in Excel Using

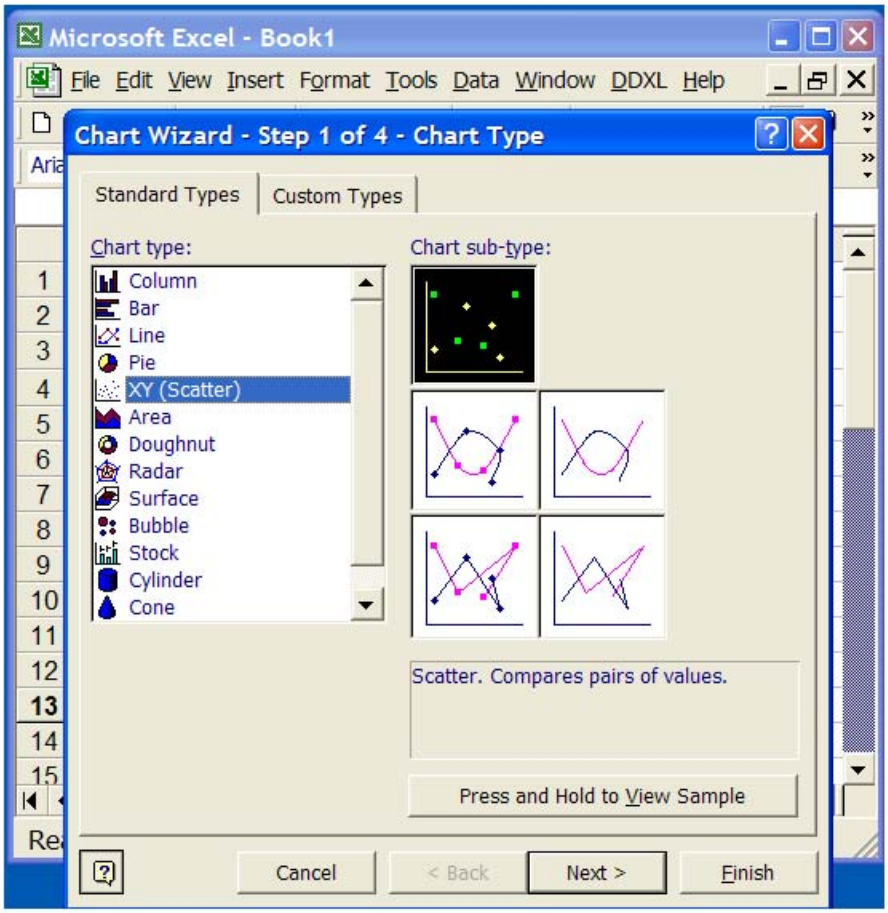
1. We will use the four step **Chart Wizard** that comes with Excel to create a scatter plot . Begin by clicking on **Insert** and then finding the **Chart…** wizard as shown on the right. Or you may find the chart wizard icon and click on it.





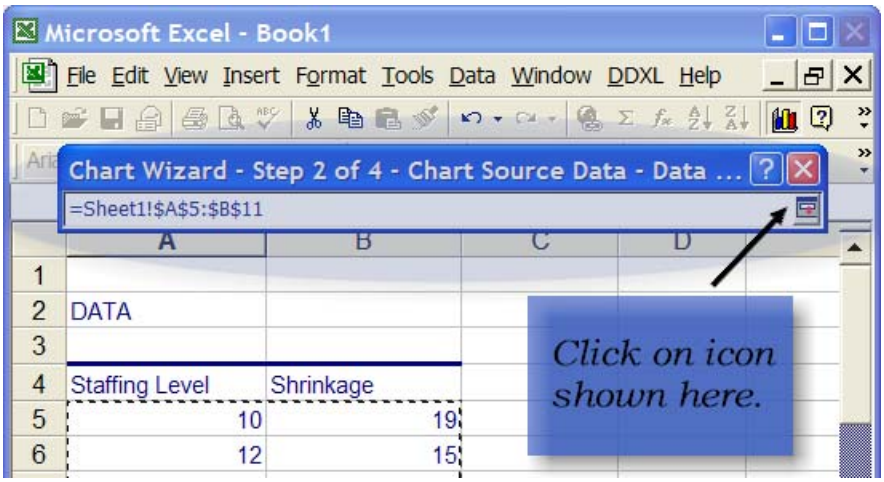
2. In **Step One** of Chart Wizard screen we will choose the **XY (Scatter)** diagram under **Chart Type**. Under the **Chart sub-type**, choose the scatter diagram without any connecting lines.

   Click on the **Next >** button to proceed to **Step Two** of the Chart Wizard. In **Step Two** we will need to define the Data Range Box. Clicking on the box as shown on the left will minimize the Chart Wizard and take you back to the worksheet.

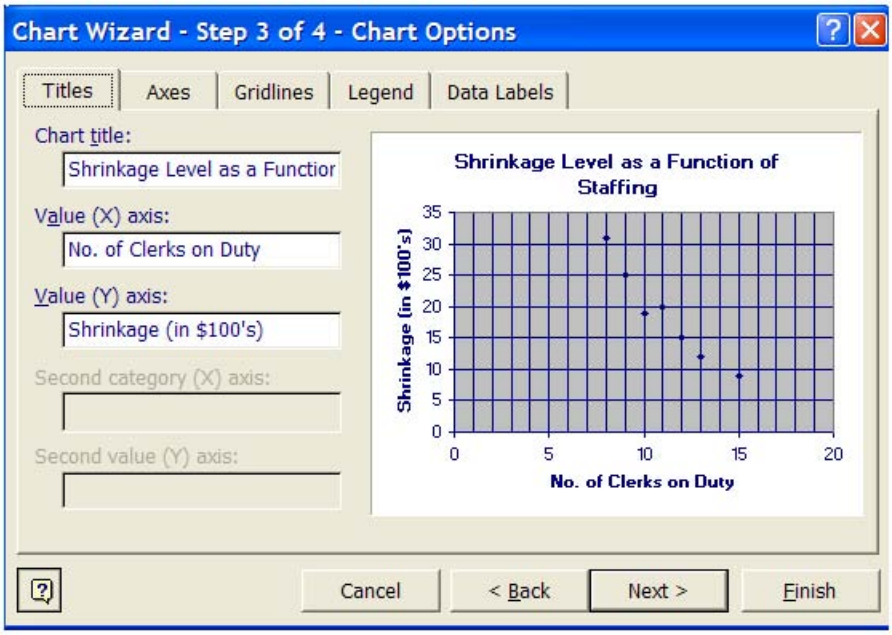


3. You will now want to *select the data range* area beginning at the upper left corner and proceeding to the lower right. When you select a group of cells, a dotted line will surround the selection. Once you have selected your data range, notice that the Chart Wizard automatically filled in this range (see the example on the right).

4. Next find the minimized Chart Wizard and click on the icon, as shown, to maximize the Chart Wizard once again. Proceed to Step Three.



5. **Step Three** allows you to fill in a title, identify the axes, choose whether or not you want grid lines, as well as other features you may explore on your own.

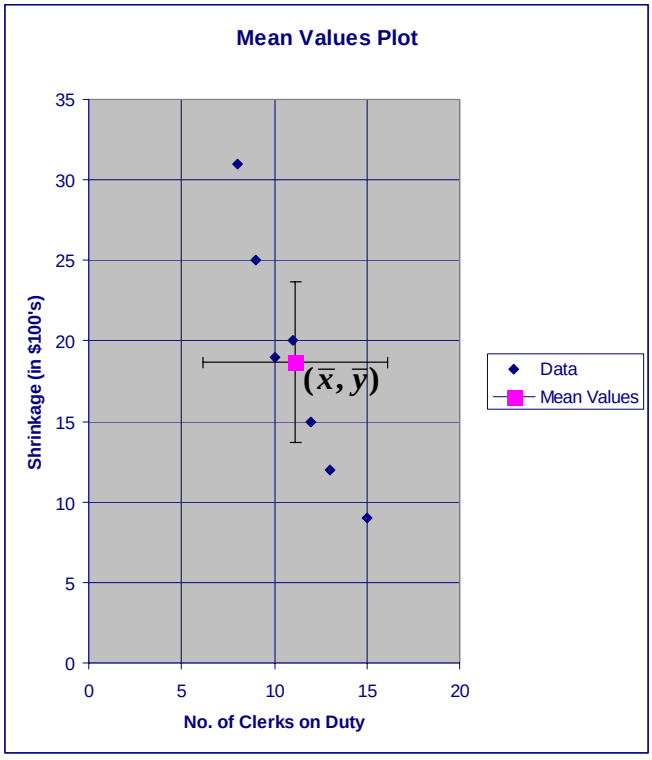


Step Four allows you to choose where you want to place the chart. Notice that once you have created the chart, you may edit it at any time by selecting the chart and right clicking inside the selected area. This gives you a range of edit options for your scatter plot.

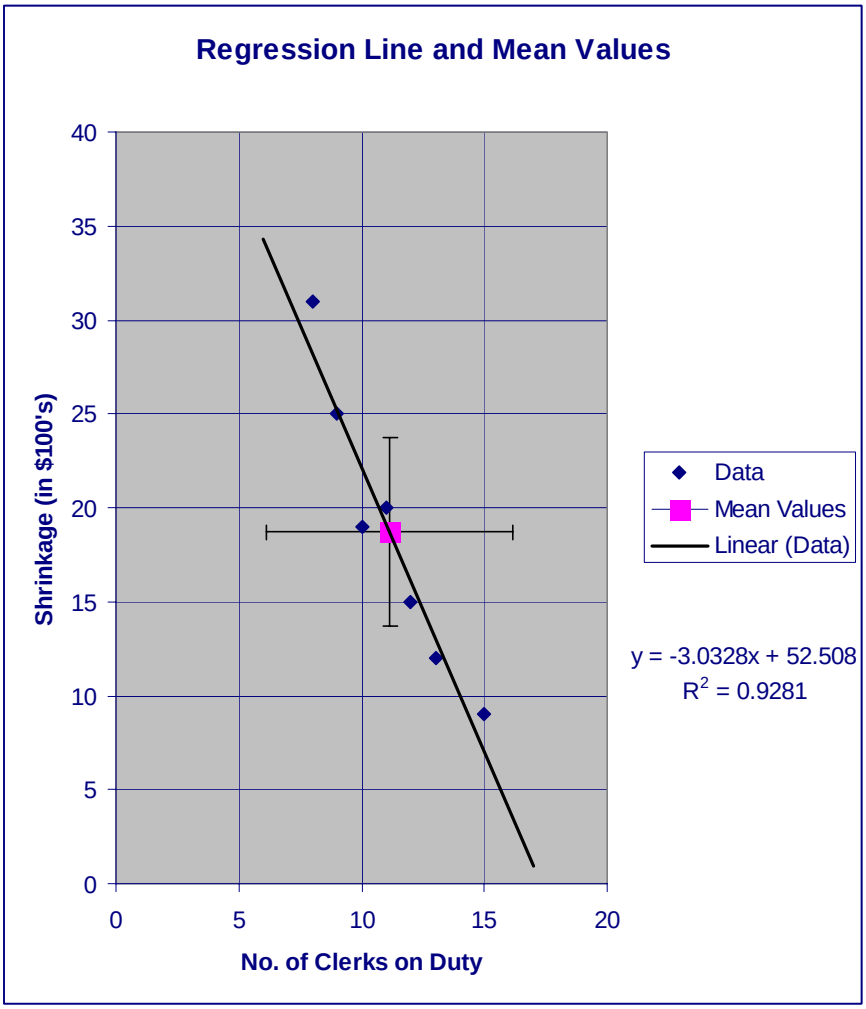
Creating a Mean Values Plot in Excel

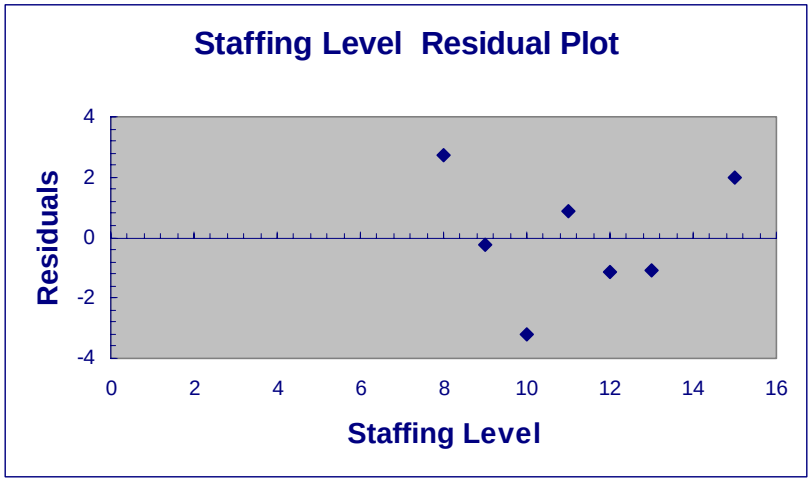
1. Once we have created our Scatter Plot, we might want to add an additional coordinate, viz. the mean values of x and y or ($\bar{x}, \bar{y}$) which in this case is the point (11.14, 18.71). Thus, the average number of clerks was 11.14, while the average dollar amount of shrinkage was $1871.

   Notice that this coordinate breaks the scatter plot into four regions. In our problem we can see that four observed values lie <u>above and to the left</u> of the mean values, and that the remainder of the data lie <u>below and to the right</u> of the mean values.



Creating a Least Squares Regression Line in Excel





Evaluating the Excel Summary Output Worksheet

DATA

Staffing Level	Shrinkage
10	19
12	15
11	20
15	9
9	25
13	12
8	31

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.963404473
R Square	0.928148178
Adjusted R Square	0.913777814
Standard Error	2.227988875
Observations	7

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	320.6088993	320.6089	64.58766
Residual	5	24.81967213	4.963934	
Total	6	345.4285714		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	52.50819672	4.288469444	12.24404	6.43E-05
Staffing Level	-3.032786885	0.377369785	-8.03664	0.000482

RESIDUAL OUTPUT

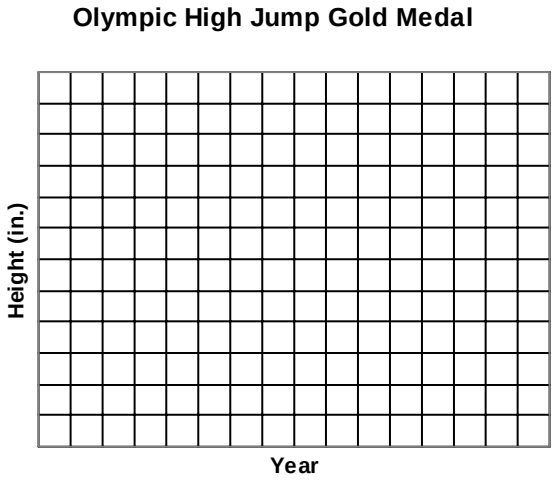
Observation	*Predicted Shrinkage*	*Residuals*
1	22.18032787	-3.18032787
2	16.1147541	-1.1147541
3	19.14754098	0.852459016
4	7.016393443	1.983606557
5	25.21311475	-0.21311475
6	13.08196721	-1.08196721
7	28.24590164	2.754098361

Linear Regression Worksheet – Olympic High Jump

Name: _____ Teacher: _____ Date: _____

A) Draw a scatterplot for the following data

Year	High Jump Height (in)
1900	74.80
1904	71.00
1908	75.00
1912	76.00
1920	76.25
1924	78.00
1928	76.38
1932	77.63
1936	79.94
1948	78.00
1952	80.32
1956	83.25
1960	85.00
1964	85.75
1968	88.25
1972	87.75
1976	88.50
1980	92.75
1984	92.50

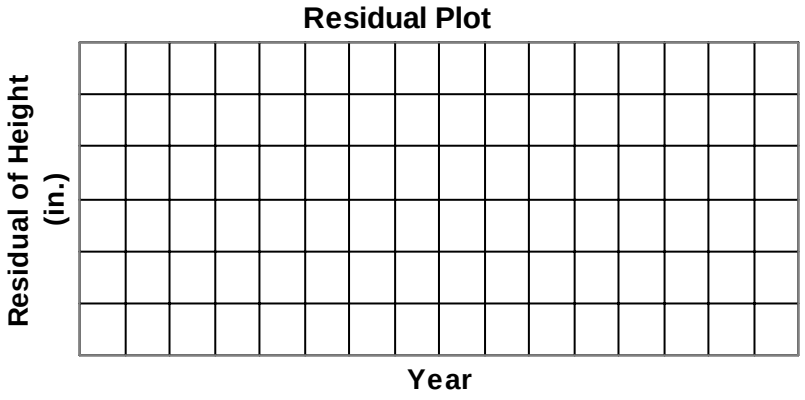


B) Find the regression equation and correlation for the Olympic High Jump data.

C) Fill out the following table.

Year	High Jump (y)	Predicted $(\hat{y})$	Residual $(y - \hat{y})$
1900	74.80		
1904	71.00		
1908	75.00		
1912	76.00		
1920	76.25		
1924	78.00		
1928	76.38		
1932	77.63		
1936	79.94		
1948	78.00		
1952	80.32		
1956	83.25		
1960	85.00		
1964	85.75		
1968	88.25		
1972	87.75		
1976	88.50		
1980	92.75		
1984	92.50		

D) Determine whether or not this model is appropriate for the given data. (Make sure that you draw a residual plot of the data.)



E) Do you need to change your model to make it a better fit? How?

F) You will notice that there is no observation for 1940 or 1944. Why not? Make a prediction for 1944. Is this a good prediction? Why or why not?

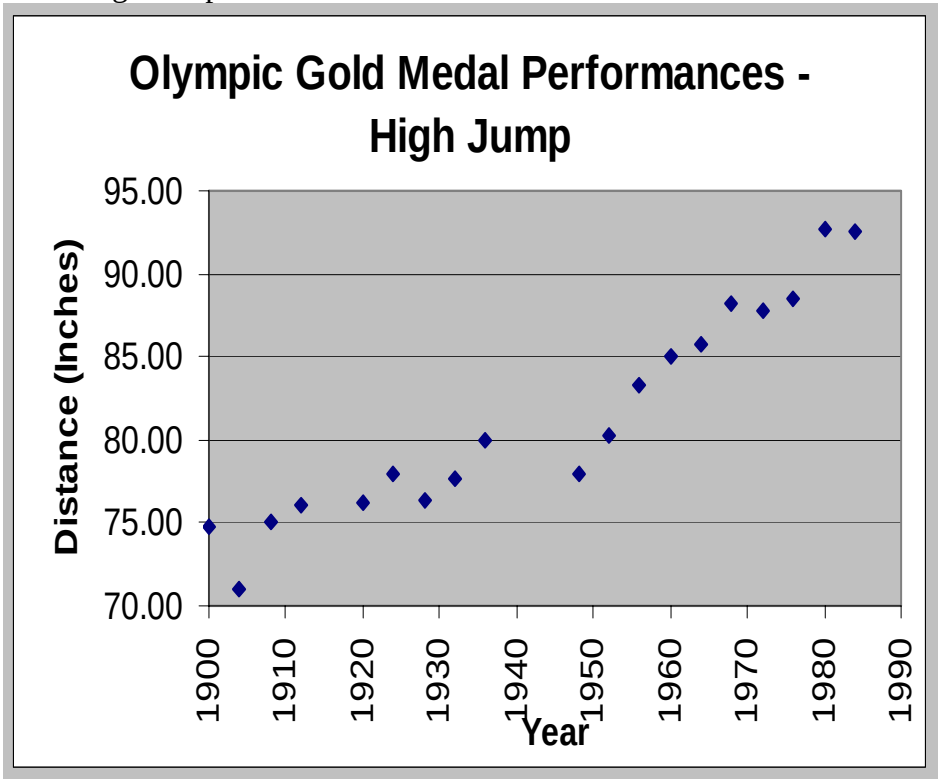
G) Can you make a prediction for the Gold Medal performance for the High Jump in the 2000 Olympics? How does that prediction compare to the actual height of 92.52 inches? What are the problems with making this prediction?

Linear Regression Worksheet – High Jump

Name: KEY Teacher: KEY Date: KEY

A) Draw a scatterplot for High Jump.

Year	High Jump
1900	74.80
1904	71.00
1908	75.00
1912	76.00
1920	76.25
1924	78.00
1928	76.38
1932	77.63
1936	79.94
1948	78.00
1952	80.32
1956	83.25
1960	85.00
1964	85.75
1968	88.25
1972	87.75
1976	88.50
1980	92.75
1984	92.50



B) Find the regression equation and correlation for the Olympic High Jump data.

Answer: $\hat{y} = -352.8792904 + 0.223479688x$ and $r = 0.953173602$ Below is the summary output of the linear regression.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.953173602
R Square	0.908539916
Adjusted R Square	0.903159911
Standard Error	1.9943117
Observations	19

ANOVA

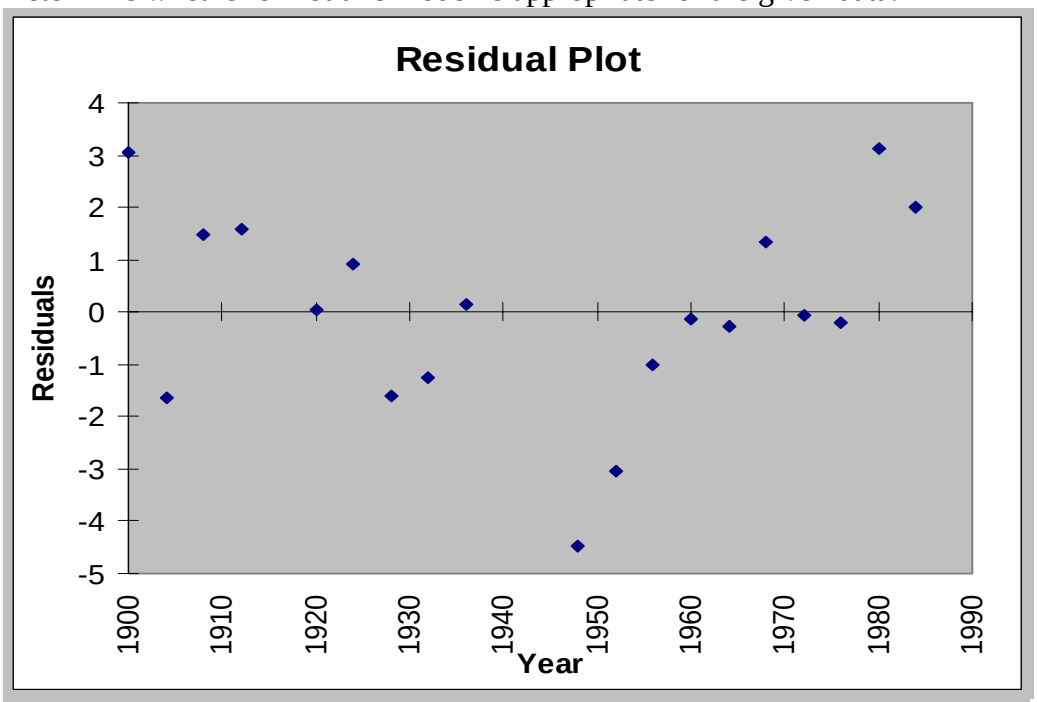
	df	*SS*	*MS*	*F*
Regression	1	671.6567945	671.65679	168.87344
Residual	17	67.61374564	3.9772792	
Total	18	739.2705401		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-352.8792904	33.4235998	-10.55779	6.947E-09
X Variable 1	0.223479688	0.017197186	12.995131	2.947E-10

C) Fill out the following table:

Year	High Jump (y)	Predicted $(\hat{y})$	Residual $(y - \hat{y})$
1900	74.80	71.73327	3.066734
1904	71.00	72.62714	-1.62714
1908	75.00	73.52101	1.478987
1912	76.00	74.41489	1.585113
1920	76.25	76.20263	0.047366
1924	78.00	77.09651	0.903493
1928	76.38	77.99038	-1.61038
1932	77.63	78.88425	-1.25425
1936	79.94	79.77813	0.161872
1948	78.00	82.45975	-4.45975
1952	80.32	83.35362	-3.03362
1956	83.25	84.2475	-0.9975
1960	85.00	85.14137	-0.14137
1964	85.75	86.03524	-0.28524
1968	88.25	86.92912	1.320884
1972	87.75	87.82299	-0.07299
1976	88.50	88.71686	-0.21686
1980	92.75	89.61074	3.139264
1984	92.50	90.50461	1.99539

D) Determine whether or not this model is appropriate for the given data.



To determine the appropriateness of the regression model, you must first look at the residual plot. Using the plot above, we see that there is a discernable pattern, which tells us that the linear model may not be appropriate.

E) Do you need to change your model to make it a better fit? How?

Answer: Looking at the residual plot and the scatterplot, there appears to be two parts to the data. In order to be able to use this model for prediction, a piece-wise model may be useful for finding the regression of the observations from 1900 to 1936 and a second regression for the observations from 1948 to 1984. When that is done, you will get $\hat{y} = -231.924 + 0.160583x$ and the following data:

SUMMARY OUTPUT

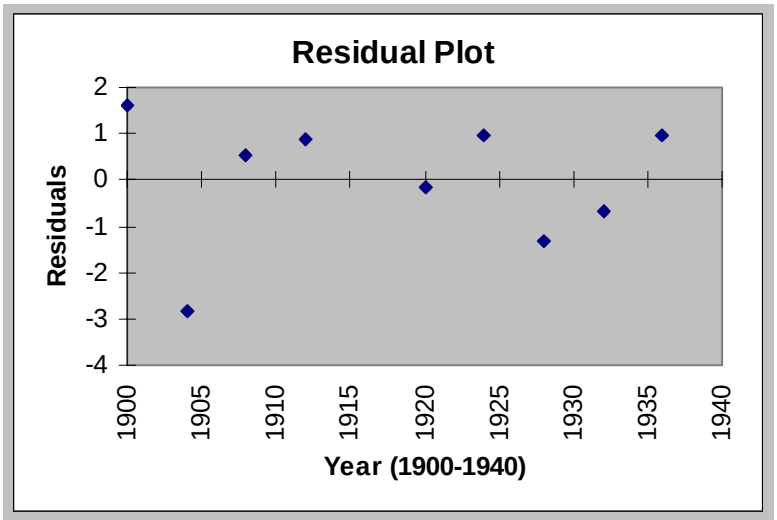
Regression Statistics

Multiple R	0.826353
R Square	0.68286
Adjusted R Square	0.637554
Standard Error	1.500253
Observations	9

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	33.924	33.924	15.07226
Residual	7	15.7553	2.250758	
Total	8	49.67931		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-231.924	79.34459	-2.92299	0.022245
X Variable 1	0.160583	0.041363	3.882301	0.006035



This residual plot shows no definitive pattern and the correlation coefficient tells us that this regression is moderately strong.

A second regression can be done on the rest of the observations from 1948-1984. When that is done, you will get $\hat{y} = -671.924 + 0.385621x$ and the following data:

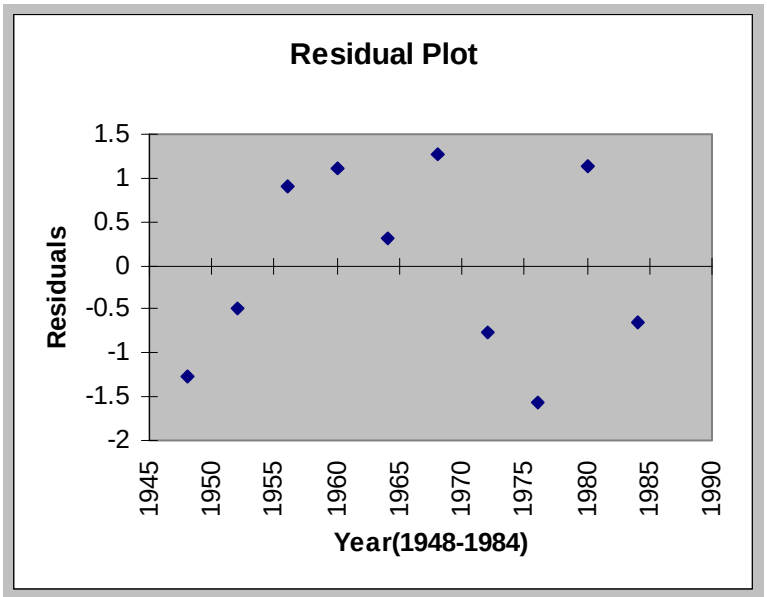
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.974617
R Square	0.949878
Adjusted R Square	0.943613
Standard Error	1.137843
Observations	10

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	196.2889	196.2889	151.611
Residual	8	10.3575	1.294688	
Total	9	206.6464		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-671.924	61.57245	-10.9127	4.41E-06
X Variable 1	0.385621	0.031318	12.31304	1.76E-06



There appears to be a curved pattern in the residual plot that may mean that a linear pattern may not be the most appropriate model. The student should be cautious when making predictions with this model.

F) You will notice that there is no observation for 1940 or 1944. Why not? Make a prediction for 1944. Is this a good prediction? Why or why not?

Answer: The Olympic were not held in 1940 or 1944 because of World War II. In making a prediction for 1944, the student needs to decide which regression that he/she should use. Two reasonable predictions are as follows.

First, take the regression for the whole data set so that the prediction will not be outside of the range of the data set. Using this model, the prediction is 81.56522307 inches. The prediction seems reasonable for all the observations together.

Second, the student could make a prediction from the 1900-1936 regression and a prediction from the 1948-1984 regression, then average the two predictions. Doing that, we get 80.249352 inches from the first regression and 77.723224 inches from the second regression. When we average the two predictions, we get 78.986288 inches, which is three inches less than the other regression prediction. This prediction also seems reasonable because there was a drop between 1936 and 1948.

G) Can you make a prediction for the Gold Medal performance for the High Jump in the 2000 Olympics? How does that prediction compare to the actual height of 92.52 inches? What are the problems with making this prediction?

Answer: Again, there are two ways of predicting when x=2000.

First, if we use the regression from the whole data set, we will get 94.0800856 inches. This is more than the actual observation. The problem with this prediction is that we are predicting outside of the data set. This is called extrapolation and such predictions cannot be trusted.

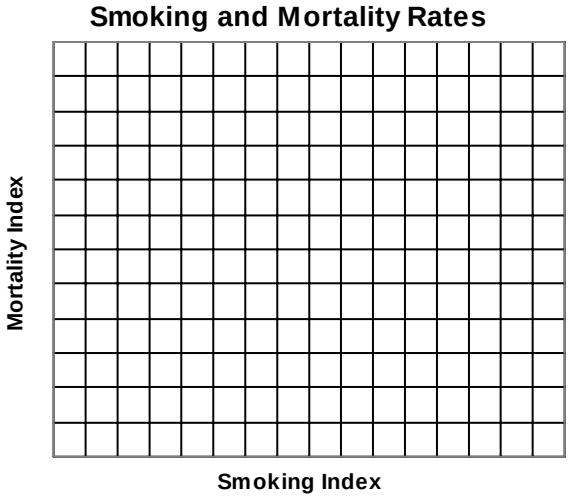
Second, we can use the 1948-1984 regression. If the student chooses this regression, the prediction is 99.318 inches. Again, this prediction is much higher than the actual value of 92.52 inches. Extrapolation cannot be trusted because it is outside of the range of the data.

Linear Regression Worksheet – Smoking Mortality

Name: _____ Teacher: _____ Date: _____

A) Draw a scatterplot of these data.

Occupation Label	Smoking Index	Mortality Index
1	77	84
2	137	116
3	117	123
4	94	128
5	116	155
6	102	101
7	111	118
8	93	113
9	88	104
10	102	88
11	91	104
12	104	129
13	107	86
14	112	96
15	113	144
16	110	139
17	125	113
18	133	146
19	115	128
20	105	115
21	87	79
22	91	85
23	100	120
24	76	60
25	66	51

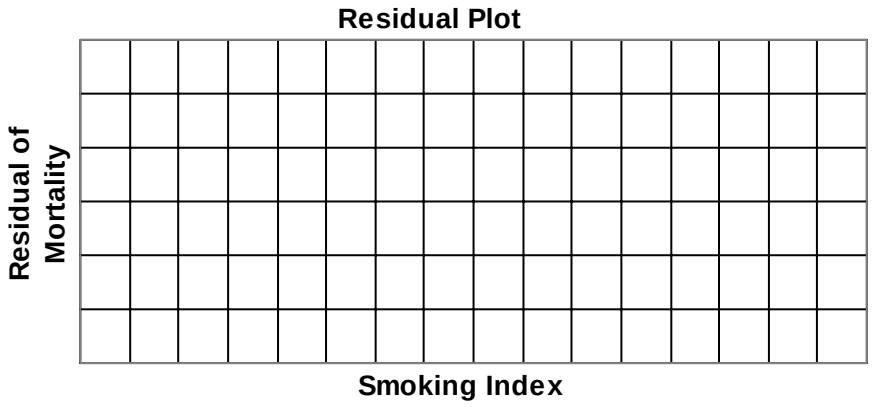


B) Find the regression equation and correlation for the Smoking and Mortality indices.

C) Fill out the following table.

Occupation Label	Smoking Index	Mortality Index (y)	Predicted Mortality $(\hat{y})$	Residual Value $(y-\hat{y})$
1	77	84		
2	137	116		
3	117	123		
4	94	128		
5	116	155		
6	102	101		
7	111	118		
8	93	113		
9	88	104		
10	102	88		
11	91	104		
12	104	129		
13	107	86		
14	112	96		
15	113	144		
16	110	139		
17	125	113		
18	133	146		
19	115	128		
20	105	115		
21	87	79		
22	91	85		
23	100	120		
24	76	60		
25	66	51		

D) Determine whether or not this model is appropriate for the given data. (Make sure that you draw a residual plot of the data.)



E) Do you need to change your model to make it a better fit? How?

F) What is the mortality for an occupation which is comparable to the national average (Smoking Index = 100)? How does this prediction compare to the actual observation?

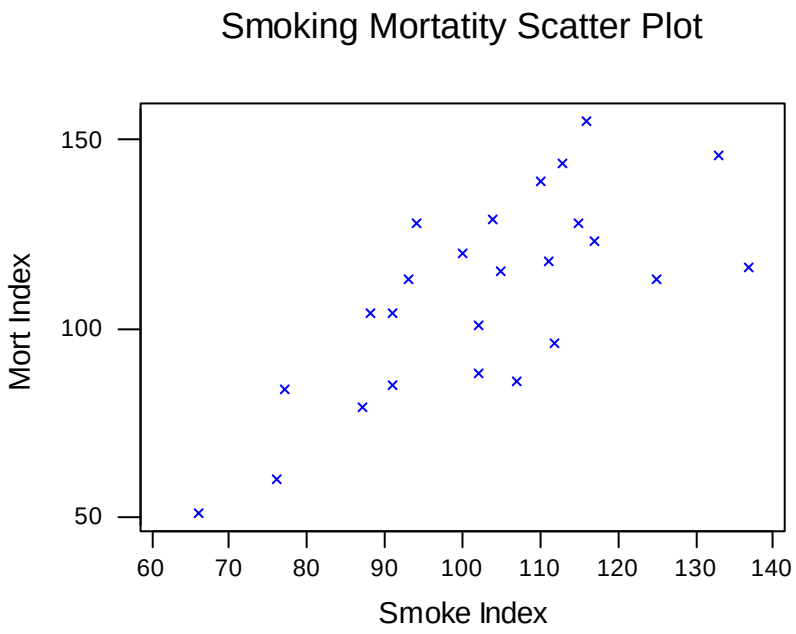
G) Can you make a prediction of an occupation in which no one smokes (Smoking Index = 0)? What problems do you have with this prediction?

Linear Regression Worksheet – Smoking Mortality

Name: KEY Teacher: KEY Date: KEY

Draw a scatterplot of these data.

Occupation Label	Smoke Index	Mort Index
1	77	84
2	137	116
3	117	123
4	94	128
5	116	155
6	102	101
7	111	118
8	93	113
9	88	104
10	102	88
11	91	104
12	104	129
13	107	86
14	112	96
15	113	144
16	110	139
17	125	113
18	133	146
19	115	128
20	105	115
21	87	79
22	91	85
23	100	120
24	76	60
25	66	51



Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Smoke In	25	102.88	104.00	103.00	17.20	3.44
Mort Ind	25	109.00	113.00	109.52	26.11	5.22

Variable	Minimum	Maximum	Q1	Q3
Smoke In	66.00	137.00	91.00	114.00
Mort Ind	51.00	155.00	87.00	128.00

B) Find the regression equation and correlation for the Smoking/Mortality data.

Answer: Using Microsoft Excel, we get a regression equation of $\hat{y} = -2.89 + 1.0875x$ with a correlation coefficient of 0.71624. The following is the printout from Miinitab.

Regression Analysis

The regression equation is
Mort Index = - 2.9 + 1.09 Smoke Index

Predictor	Coef	StDev	T	P
Constant	-2.89	23.03	-0.13	0.901
Smoke In	1.0875	0.2209	4.92	0.000

S = 18.62 R-Sq = 51.3% R-Sq(adj) = 49.2%

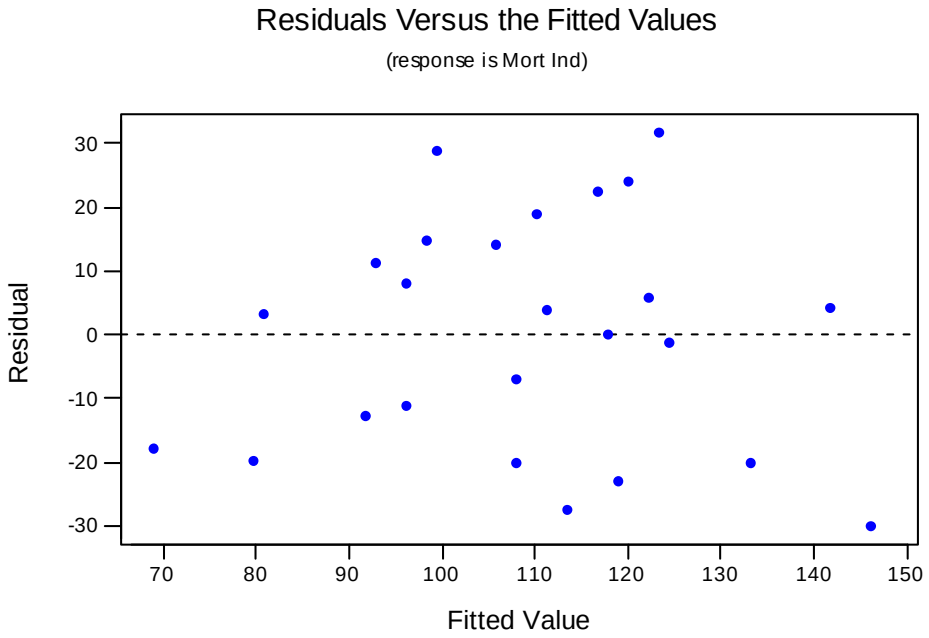
C) Fill out the following table.

Occupation Label	Smoking Index	Mortality Index (y)	Predicted Mortality $(\hat{y})$	Residual Value $(y - \hat{y})$
1	77	84	80.85467	3.145335
2	137	116	146.1066	-30.1066
3	117	123	124.356	-1.35596
4	94	128	99.34271	28.65729
5	116	155	123.2684	31.73158
6	102	101	108.043	-7.04297
7	111	118	117.8308	0.169238
8	93	113	98.25518	14.74482
9	88	104	92.81752	11.18248
10	102	88	108.043	-20.043
11	91	104	96.08012	7.919883
12	104	129	110.218	18.78196
13	107	86	113.4806	-27.4806

14	112	96	118.9183	-22.9183
15	113	144	120.0058	23.99417
16	110	139	116.7432	22.25677
17	125	113	133.0562	-20.0562
18	133	146	141.7565	4.243528
19	115	128	122.1809	5.819109
20	105	115	111.3056	3.694432
21	87	79	91.72999	-12.73
22	91	85	96.08012	-11.0801
23	100	120	105.8679	14.13209
24	76	60	79.76713	-19.7671
25	66	51	68.89181	-17.8918

D) Determine whether or not this model is appropriate for the given data. (Make sure that you draw a residual plot of the data.)

Answer: The residual plot to the left is a very good indication that the regression line is a good fit. The plot does not have a discernable pattern, and the correlation coefficient and coefficient of determination are reasonable.



E) Do you need to change your model to make it a better fit? How?
Answer: There is no need to make any changes to make it a better fit.

F) What is the mortality for an occupation which is comparable to the national average (Smoking Index = 100)? How does this prediction compare to the actual observation?

Answer: If the student is given 100 for the Smoking Index, he/she should plug it into the regression equation which would give an answer of 105.64668, which is a reasonable answer. It is almost 15 points lower than the actual observation, but it is within all of the other points in the data.

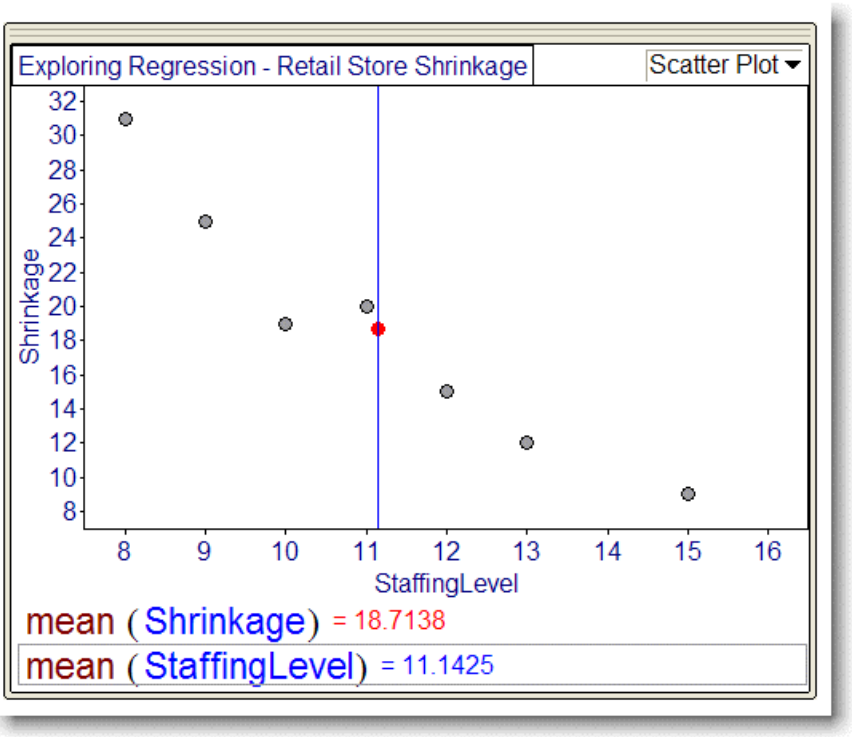
G) Can you make a prediction of an occupation in which no one smokes (Smoking Index = 0)? What problems do you have with this prediction?

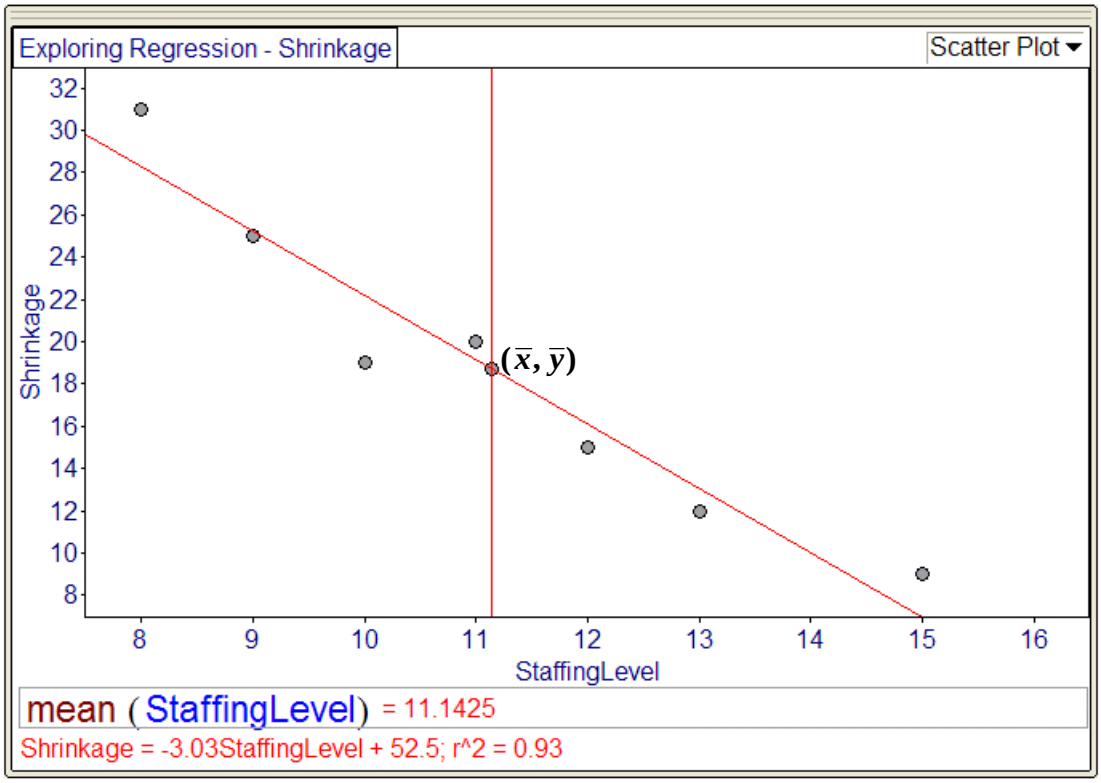
Answer: For a Smoking Index of 0, the Mortality Index value would be -2.88532. This is unrealistic, because we cannot assume that every worker in one occupation does not smoke. Also, the student is being asked to make a prediction outside of the data range, which can never be trusted.

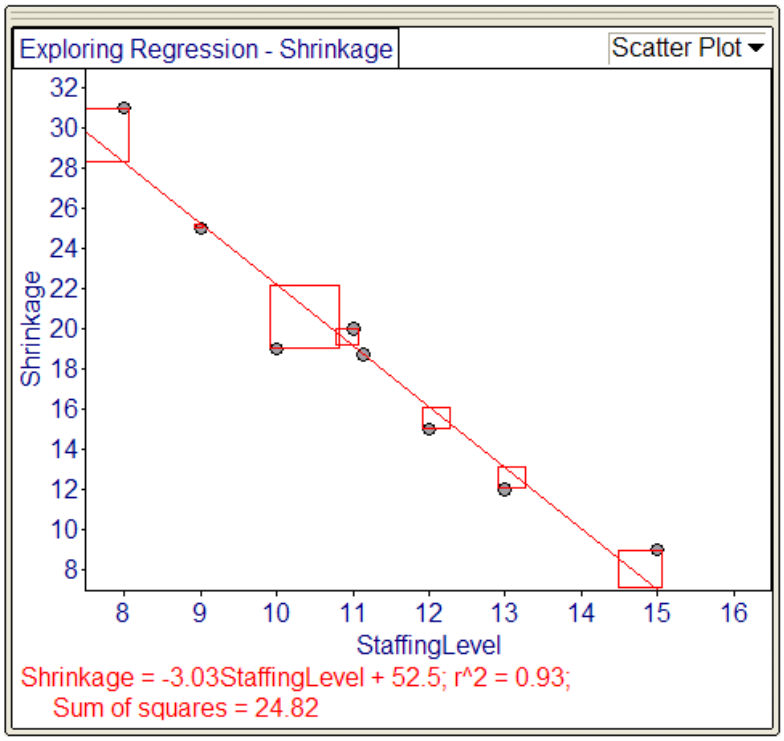
Part IV. Using Fathom in Linear Regression Analysis

Exploring Regression - Shrinkage

	StaffingLevel	Shrinkage
1	10	19
2	12	15
3	11	20
4	15	9
5	9	25
6	13	12
7	8	31
8	11.14	18.71







Exploring Regression - Shrinkage

	StaffingLevel	Shrinkage	Predicted	Residual
1	10	19	22.1787	-3.17871
2	12	15	16.1131	-1.11314
3	11	20	19.1459	0.854078
4	15	9	7.01478	1.98522
5	9	25	25.2115	-0.211494
6	13	12	13.0804	-1.08035
7	8	31	28.2443	2.75572
8	11.14	18.71	18.7213	-0.011332